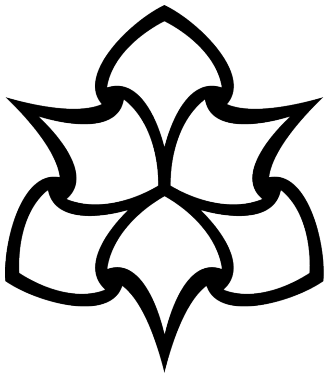

Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP

Angelo Moretti[1], Natalie Shlomo[2] and Joseph W. Sakshaug[3]

1 (Corresponding author). University of Sheffield, Department of Geography. Email: a.moretti@sheffield.ac.uk or a.moretti2@outlook.com

2 University of Manchester, Social Statistics Department. Email: natalie.shlomo@manchester.ac.uk

3 German Institute for Employment Research, Nuremberg. Email: joe.sakshaug@iab.de

Parametric Bootstrap Mean Squared Error of a Small Area Multivariate EBLUP

Manuscript

Abstract

This article deals with mean squared error (MSE) estimation of a multivariate empirical best linear unbiased predictor (MEBLUP) under the unit-level multivariate nested-errors regression model for small area estimation via parametric bootstrap. A simulation study is designed to evaluate the performance of our algorithm and compare it with the univariate case bootstrap MSE which has been shown to be consistent to the true MSE. The simulation shows that, in line with the literature, MEBLUP provides unbiased estimates with lower MSE than EBLUP. We also provide a short empirical analysis based on real data collected from the U.S. Department of Agriculture.

Keywords: Multivariate empirical best linear unbiased predictor; Model-based inference; Multivariate small area estimation; Multivariate multilevel models; Resampling

1. Introduction

Regional policies need to base their funding allocation on reliable statistical information. However, large-scale social sample surveys are typically not designed to be representative at a low geographical level. Thus, small area estimation (SAE) methods based on models might provide more accurate estimates than direct estimators (Rao and Molina, 2015). Mixed effects linear regression models are traditionally used in order to provide more accurate estimates than design-based estimation techniques. These kinds of models have been used extensively in the literature, and for a detailed review of these in SAE we refer to Rao and Molina (2015). Estimating the precision of small area estimates is a crucial and challenging exercise (Marchetti et al., 2012).

As Molina (2009) points out, when the target of inferential interest is a random vector, multivariate regression models might be a natural model setting. Indeed, multivariate models take into account the correlation structure among the vector components; hence, it is possible to improve the precision of the estimates over the univariate case (Datta et al., 1999). Fuller and Harter (1987) develop a multivariate mixed effects model to predict a vector of means of multiple characteristics of a finite population. Datta et al. (1999) propose a multivariate empirical best linear unbiased predictor (MEBLUP) and empirical bayes (EB) approach for small area mean vectors along with an approximation for the mean squared error (MSE). Some recent work in the literature are Molina (2009) and Baillo and Molina (2009). Molina (2009) deals with the multivariate mixed effects model with logarithmic transformation, and Baillo and Molina (2009) study a particular case of the multivariate nested error regression model with correlated sampling errors. Both papers provide analytical approximations for the MSE.

The best linear unbiased predictor (BLUP) depends on unknown quantities (variance components). When these quantities are estimated using suitable estimation techniques, we obtain the empirical BLUP (EBLUP). Unfortunately, the exact MSE of an EBLUP cannot be obtained in closed form; therefore, approximations have been proposed in the literature (González-Manteiga et al., 2008a). Kackar and Harville (1981) propose an approximation of the MSE assuming normality of the errors and random effects. Prasad and Rao (1990) obtain an MSE approximation for models with block-diagonal covariance matrices. Datta and Lahiri (2000) provide analytical approximations for general models with a block-diagonal structure when variance components are estimated by maximum likelihood (ML) or restricted maximum likelihood (REML). Das et al. (2004) deal with approximations for a wider class of models. In multivariate SAE, Datta et al. (1999) propose a second-order unbiased analytical approximation for the MSE of a multivariate EBLUP following Datta and Lahiri (2000).

When the MSE exact analytical estimator cannot be computed, an alternative way to approximate the MSE is via bootstrap techniques. It is important to highlight that, even when large sample approximations are available, the bootstrap may provide more accurate estimation alternatives due to its second-order accuracy (González-Manteiga et al., 2008a). This property is not achieved by the majority of asymptotic methods. We refer to Efron and Tibshirani (1993) and Hall (1992) for a broader discussion of this property.

In this article, we assume that the values of the target vector in the units of a finite population are realizations of a random multivariate variable following the Fuller and Harter multivariate mixed effects model (Fuller and Harter, 1987). We propose a maximum likelihood (ML)-based parametric bootstrap procedure designed for estimating a vector of MSEs for a vector of means when the auxiliary variables are available at the unit-level.

This paper is organised as follows. In section 2 the multivariate mixed effects model is reviewed along with the multivariate EBLUP. In section 3 we discuss the MSE estimation via parametric bootstrap. In section 4 we study the behaviour of our bootstrap MSE in a model-based simulation study and compare it with the univariate case. In section 5 we present an example based on survey data on corn and soy bean production. In section 6 we conclude with some final remarks.

2. Multivariate Small Area Estimation of a Means Vector

Let $d = 1, \dots, D$ denote the small areas for which we want to compute the estimates and let us consider a sample $s \subset \Omega$ of size n drawn from the target finite population Ω of size N. The non-sampled units, $N - n$ are denoted by r, hence, $s_d = s \cap \Omega_d$ is the sub-sample from the small area d of size n_d, $n = \sum_{d=1}^{D} n_d$, and $s = \cup_d s_d$. r_d denotes the non-sampled units for small area d of $N_d - n_d$ dimension.

Considering $\boldsymbol{y}_{di} = (y_{di1}, \dots, y_{diK})$, which denotes the K-dimensional row vector of observations on the target K variables for $i = 1, \dots, N_d$ and $d = 1, \dots, D$, we can define the target mean vector as follows:

$$\bar{\boldsymbol{y}}_d = N_d^{-1} \sum_{i=1}^{N_d} \boldsymbol{y}_{di}. \tag{1}$$

Because of linearity of this quantity, each area mean vector can be split into sampled and non-sampled (out-of-sample) elements as follows:

$$\bar{\boldsymbol{y}}_d = N_d^{-1}\left(\sum_{i\in s_d} \boldsymbol{y}_{di} + \sum_{i\in r_d} \boldsymbol{y}_{di}\right). \tag{2}$$

The quantity $\sum_{i\in r_d} \boldsymbol{y}_{di}$ is not observed, so it needs to be predicted. In this article we make use of the multivariate mixed effects model advocated in unit-level SAE by Fuller and Harter (1987).

2.1 Multivariate nested-error linear regression model

We assume that the following linear model relates the response variables to the covariates in the population as follows (Fuller and Harter, 1987):

$$\xi: \boldsymbol{y}_{di} = \boldsymbol{x}_{di}\boldsymbol{\beta} + \boldsymbol{u}_d + \boldsymbol{e}_{di}, \qquad d = 1, \dots, D, i = 1, \dots, N_d, \tag{3}$$

$$\boldsymbol{u}_d \overset{iid}{\sim} N_K(\boldsymbol{0}, \boldsymbol{\Sigma}_u), \qquad \boldsymbol{e}_{di} \overset{iid}{\sim} N_K(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$$

where $\boldsymbol{x}_{di}$ is a p-dimensional row vector of auxiliary variables, $\boldsymbol{\beta}$ is a $p \times K$ matrix of unknown regression coefficients, $\boldsymbol{u}_d$ is a K-dimensional row vector of area effects, and $\boldsymbol{e}_{di}$ is K-dimensional row vector of the individual effects; $\boldsymbol{u}_d$ and $\boldsymbol{e}_{di}$ are assumed to be independent and normally distributed, N_K denotes a K-variate Normal distribution. Here, the $K \times K$ positive-definite matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$ are the variance-covariance matrices of the area effects and individual effects, respectively.

Under model (3) we can write the realised mean of area d as:

$$\bar{\boldsymbol{y}}_d = \bar{\boldsymbol{x}}_{d,pop}\boldsymbol{\beta} + \boldsymbol{u}_d \tag{4}$$

where $\bar{\boldsymbol{x}}_{d,pop}$ denotes the known population means of $\boldsymbol{x}_{di}$ for area d.

2.2 Estimation and prediction of unknown parameters

For simplicity we now make use of the following notation (Fuller and Harter, 1987):

$\boldsymbol{Y}' = (\boldsymbol{y}_{11}, \boldsymbol{y}_{12}, \ldots, \boldsymbol{y}_{1,n_1}, \ldots, \boldsymbol{y}_{D1}, \ldots, \boldsymbol{y}_{D,n_D})$,

$\boldsymbol{X}' = [(\boldsymbol{I}_K \otimes \boldsymbol{x}_{11})', (\boldsymbol{I}_K \otimes \boldsymbol{x}_{12})', \ldots, (\boldsymbol{I}_K \otimes \boldsymbol{x}_{1,n_1})', \ldots, (\boldsymbol{I}_K \otimes \boldsymbol{x}_{D,n_D})']$,

where $\boldsymbol{Y}$ denotes the vector of NK observations on $\boldsymbol{y}_{di}$ where $\boldsymbol{y}_{di}$ is defined above, and $\boldsymbol{X}$ denotes the $NK \times pK$ matrix of covariates. The operator $\otimes$ denotes the Kronecker product, and $\boldsymbol{I}$ denotes the identity matrix.

Let us now denote the covariance matrix of $\boldsymbol{Y}$ by

$$\boldsymbol{V}(\boldsymbol{Y}) = block\ diag(\boldsymbol{V}_{11}, \ldots, \boldsymbol{V}_{DD}) \tag{5}$$

where $\boldsymbol{V}_{dd} = (\boldsymbol{J}_{dd} \otimes \boldsymbol{\Sigma}_u) + \left(\boldsymbol{I}_{n_d} \otimes \boldsymbol{\Sigma}_e\right)$. $\boldsymbol{J}_{dd}$ is the $n_d \times n_d$ matrix with every element equal to one. Let $vec\ \boldsymbol{\beta}$ denote the column vector of dimension pK obtained by listing the columns of $\boldsymbol{\beta}$ one under the other starting from the first column. The estimator of $vec\ \boldsymbol{\beta}$ is:

$$vec\ \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}. \tag{6}$$

The empirical best linear unbiased predictors of the random effects are given by the following expression (Fuller and Harter, 1987):

$$\widehat{\boldsymbol{u}}_d = (\overline{\boldsymbol{y}}_{d,s} - \overline{\boldsymbol{x}}_{d,s}\widehat{\boldsymbol{\beta}})[\left(\widehat{\boldsymbol{\Sigma}}_u + n_d^{-1}\widehat{\boldsymbol{\Sigma}}_e\right)^{-1}\widehat{\boldsymbol{\Sigma}}_u], \qquad d = 1, \dots, D \tag{7}$$

where $\overline{\boldsymbol{y}}_{d,s}$ denotes the sample mean vector and $\overline{\boldsymbol{x}}_{d,s}$ denotes the means of the auxiliary variables in area d. The index 's' refers to the sample quantities. $\widehat{\boldsymbol{\Sigma}}_u$ and $\widehat{\boldsymbol{\Sigma}}_e$ are estimators of $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_e$, respectively. We refer to Schafer et al. (2002) for the estimation algorithm where the maximum likelihood (ML) approach is used.

The Multivariate Empirical Best Linear Unbiased Predictor (MEBLUP) of $\overline{\boldsymbol{y}}_d$ is given by (Fuller and Harter, 1987; Rao and Molina, 2015):

$$\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP} = \overline{\boldsymbol{x}}_{d,pop}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}_d, \qquad d = 1, \dots, D \tag{8}$$

where $\overline{\boldsymbol{x}}_{d,pop}$ denotes the known population means vector.

3. Parametric Bootstrap

This section introduces a bootstrap algorithm approximation of the MSE of $\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP}$ denoted by $\boldsymbol{MSE}\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP}\right)$ and given by the following (Kackar and Harville, 1984):

$$\boldsymbol{MSE}\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP}\right) = E\left[\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP} - \overline{\boldsymbol{y}}_d\right)\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP} - \boldsymbol{y}_d\right)'\right] = \tag{9}$$

$$= \boldsymbol{MSE}\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MBLUP}\right) + E\left[\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP} - \hat{\bar{\boldsymbol{y}}}_d^{MBLUP}\right)\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP} - \hat{\bar{\boldsymbol{y}}}_d^{MBLUP}\right)'\right] +$$

$$+E\left[\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP}-\hat{\bar{\boldsymbol{y}}}_d^{MBLUP}\right)\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP}-\bar{\boldsymbol{y}}_d\right)'\right]$$

$$+E\left[\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MBLUP}-\bar{\boldsymbol{y}}_d\right)\left(\hat{\bar{\boldsymbol{Y}}}_d^{MEBLUP}-\hat{\bar{\boldsymbol{y}}}_d^{MBLUP}\right)'\right].$$

where we denote the Multivariate Best Linear Unbiased Predictor of $\bar{\boldsymbol{y}}_d$ (assuming known covariance matrices) by $\hat{\bar{\boldsymbol{y}}}_d^{MBLUP}$. It can be shown that the last two terms of equation (9) are equal to zero for any unbiased and translation invariant estimator of the variance components (Kackar and Harville, 1984). The term $E\left[\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP}-\hat{\bar{\boldsymbol{y}}}_d^{MBLUP}\right)\left(\hat{\bar{\boldsymbol{y}}}_d^{\,MEBLUP}-\hat{\bar{\boldsymbol{y}}}_d^{MBLUP}\right)'\right]$ accounts for the estimation of the variance components.

We propose to use the parametric bootstrap procedure proposed by González-Manteiga et al. (2008a) extended to the multivariate mixed effects model used in this article. Let Ω be a finite population of dimension N generated by the superpopulation model given by (3), and let $\bar{\boldsymbol{y}}_d = N_d^{-1}\sum_{i=1}^{N_d}\boldsymbol{y}_{di}$ be the linear vector of target parameters of Ω**.** Let *s* be a random sample drawn from Ω of dimension n, using a specific sampling design.

We list the steps of the algorithm as follows:

1. Fit the multivariate model (3) to the sample *s,* $\boldsymbol{y}_s = (\boldsymbol{y}'_{1s},\dots,\boldsymbol{y}'_{Ds})'$, and obtain the estimates of the model parameters: let us denote the estimates as $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_u$, and $\hat{\boldsymbol{\Sigma}}_e$.

2. Generate the bootstrap area effects $\boldsymbol{u}_d^{*(b)} \overset{iid}{\sim} N_K\left(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_u\right), d = 1,\dots,D$. We use the symbol * for the bootstrap quantities, while (b) refers to the index of the b^{th} bootstrap replication, $b = 1,\dots,B$.

3. Generate the bootstrap errors for the sample units $\boldsymbol{e}_{di}^{*(b)} \overset{iid}{\sim} N_K\left(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_e\right), i \in s_d$ independently

of the $\boldsymbol{u}_d^{*(b)}$, $d = 1, \ldots, D$.

4. Calculate the true means vectors for each small area of the bootstrap population as follows:

$$\overline{\boldsymbol{y}}_d^{*(b)} = \overline{\boldsymbol{x}}_{d,pop}\widehat{\boldsymbol{\beta}} + \boldsymbol{u}_d^{*(b)}, d = 1, \ldots, D, \tag{10}$$

where $\overline{\boldsymbol{x}}_{d,pop}$ denotes the means of the known population auxiliary variables.

5. Generate the responses for the sample units by using the sample covariates vectors $\boldsymbol{x}_{di}$, $i \in s_d$:

$$\xi^*: \boldsymbol{y}_{di}^{*(b)} = \boldsymbol{x}_{di}\widehat{\boldsymbol{\beta}} + \boldsymbol{u}_d^{*(b)} + \boldsymbol{e}_{di}^{*(b)}, d = 1, \ldots, D, \tag{11}$$

The bootstrap sample data vector is denoted by $\boldsymbol{y}_s^{*(b)} = \left[\left(\boldsymbol{y}_{1s}^{*(b)}\right)', \ldots, \left(\boldsymbol{y}_{Ds}^{*(b)}\right)'\right]'$. Under model ξ^*, given s, the MSE of $\overline{\boldsymbol{y}}_d^{MEBLUP*}$ is denoted by $\boldsymbol{MSE}_*(\widehat{\overline{\boldsymbol{y}}}_d^{MEBLUP*})$. Hence, for estimating the MSE of $\widehat{\overline{\boldsymbol{y}}}_d^{MEBLUP}$ given in (9), we propose to use the bootstrap MSE.

6. Fit model (3) to the bootstrap sample data $\boldsymbol{y}_s^{*(b)}$ and obtain the bootstrap MEBLUPs $\widehat{\overline{\boldsymbol{y}}}_d^{*(b)}, d = 1, \ldots, D$.

7. Replicate steps (2) through (6) for $b = ,1 \ldots, B$. The Monte Carlo approximation of the bootstrap estimator $\boldsymbol{MSE}_*(\widehat{\overline{\boldsymbol{y}}}_d^{MEBLUP*})$ is given by:

$$\boldsymbol{mse}_*(\widehat{\bar{\boldsymbol{y}}}_d^{MEBLUP*}) = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\bar{\boldsymbol{y}}}_d^{*(b)} - \bar{\boldsymbol{y}}_d^{*(b)}\right)\left(\widehat{\bar{\boldsymbol{y}}}_d^{*(b)} - \bar{\boldsymbol{y}}_d^{*(b)}\right)', d = 1, \dots, D. \quad (12)$$

We note that when $B \to \infty$, $\boldsymbol{mse}_*(\widehat{\bar{\boldsymbol{y}}}_d^{MEBLUP*})$ is a consistent estimator of $\boldsymbol{MSE}_*(\widehat{\bar{\boldsymbol{y}}}_d^{MEBLUP*})$ (Rao and Molina, 2015).

The parametric bootstrap procedure has been proven to be consistent as an estimator of the true MSE under the univariate unit-level model (González-Manteiga et al., 2008a) and the Fay-Herriot model (González-Manteiga et al., 2008b). In general, the proofs in these papers have been based on the fact that the final estimate of the MSE obtained by the bootstrap procedure is consistent if the model parameter estimates are consistent. Since we are using the Maximum Likelihood estimators for estimating the model parameters in the multivariate SAE approach, which have well-known consistency properties as shown in Sweeting (1980) and Mardia and Marshall (1984), we can prove the consistency of our proposed parametric bootstrap algorithm to the true MSE by the method of imitation.

4. Simulation Study

This simulation is designed to study the performance of the bootstrap MSE estimator presented in Section 3 under a multivariate mixed effects model when the target vector parameter is a vector of means. The results are compared with the "truth" as described in Section 4.2 and the aim is to show that a multivariate bootstrap procedure will be appropriate in the case of multivariate SAE. The bias is also studied. In the case of the univariate SAE, the bootstrap MSEs are compared with the Prasad-Rao analytical approximation of MSE (Prasad and Rao, 1990). Software developed by Yucel (2010) and Molina and Marhuenda

(2015) are used in order to estimate parameters of the multivariate and univariate models, respectively. We list the details of the functions in the appendix.

4.1 Generating the population

The simulation is a model-based simulation, where $S = 1{,}000$ populations are generated, then a sample from each population is extracted. We generate an unbalanced population using parameters with $N = 20{,}000$, $D = 80$, and $130 \leq N_d \leq 420$. N_d, $d = 1, \dots, D$ is generated from the discrete Uniform distribution, $N_d \sim dUnif(130, 420)$, with $\sum_{d=1}^{D} N_d = 20{,}000$.

The simulation modelling parameters have been chosen according to survey and satellite data for corn and soy beans in 12 Iowa counties, obtained from the 1978 June survey of the U.S. Department of Agriculture and from land observatory satellites, also known as LANDSAT during the 1978 growing season. These data were also used by Datta et al. (1999).

$\boldsymbol{y}_{di}$ observations are generated according to the multivariate mixed effects model (3) described in section 2. Here we consider a bivariate model with $k = 1{,}2; K = 2$. In this section, we use the following notation, Y_k for $k = 1{,}2$ which denote the target variables. Regarding the auxiliary variables, we have $(p = 3)$ $\boldsymbol{x}_{di} = (1, x_{di1}, x_{di2})$. The two uncorrelated covariates are generated from the discrete Uniform distribution as follows:

$$x_{di1} \sim dUnif(145, 459), \qquad x_{di2} \sim dUnif(55, 345).$$

The regression coefficients are given in the following matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} 17.97 & 0.36 & -0.03 \\ -16.35 & 0.02 & 0.50 \end{bmatrix}$$

The variance-covariance matrices are given by:

$$\boldsymbol{\Sigma}_e = \begin{bmatrix} 297.71 & -150.82 \\ -150.82 & 170.29 \end{bmatrix}$$

$$\boldsymbol{\Sigma}_u = \begin{bmatrix} 63.31 & 35.35 \\ 35.35 & 219.32 \end{bmatrix}$$

with associated correlation coefficients $\rho_e = -0.7$ and $\rho_u = 0.3$, respectively. The intra-class correlations are 0.2 and 0.6 for the first and second components, respectively; these have been chosen according to the LANDSAT data. We also studied the case where ρ_e and ρ_u have the same signs i.e. $\rho_e = 0.7$ and $\rho_u = 0.3$.

For computational reasons, we did not perform a simulation study varying many ρ_e, ρ_u and intra-class correlation coefficient values. For more details on the role of these in multivariate SAE we refer to Datta et al. (1999). Their paper shows that when ρ_e and ρ_u have opposite signs the multivariate modelling performs much better than the univariate modelling in terms of MSE. Of course, when ρ_e and ρ_u are small (theoretically tending to zero), we are close to the independence case, where the univariate modelling performs identically to the multivariate modelling Datta et al. (1999).

The steps of the simulation are as follows, for $d = 1, \dots, D$ and $s = 1, \dots, S$, with $S = 1000$:

1. *Populations generation:* generate $\boldsymbol{y}_{dis}$ according to model (3) for $s = 1, \dots, S$, with parameters presented above;
2. *Sample selection:* draw a simple random sample without replacement of size $n = 1{,}000$ from each simulated population;
3. Fit the univariate Battese, Harter and Fuller model (BHF) (Battese et al., 1988) on each sample *s* and obtain the estimates of the model parameters: $\hat{\sigma}_{es}^2, \hat{\sigma}_{us}^2$ and $\widehat{\boldsymbol{\beta}}_s^{BHF}$, thus the univariate EBLUPs are estimated: $\hat{\bar{y}}_{ds,1}^{EBLUP}$ and $\hat{\bar{y}}_{ds,2}^{EBLUP}$;
4. Estimate the MSEs of $\hat{\bar{y}}_{ds,1}^{EBLUP}$ and $\hat{\bar{y}}_{ds,2}^{EBLUP}$ on each sample *s* via parametric bootstrap (González-Manteiga et al., 2008a) with $B = 500$ replications and Prasad-Rao analytical

approximation (PR) according to Prasad and Rao (1999). In the economy of space, the PR approximation is estimated for the $\rho_e = -0.7$ and $\rho_u = 0.3$ case only;

5. Fit the multivariate mixed effects model given in (3) on each sample *s* and obtain the model parameters estimates: $\widehat{\boldsymbol{\Sigma}}_{es}, \widehat{\boldsymbol{\Sigma}}_{us}$ and $\widehat{\boldsymbol{\beta}}_s$, and the multivariate EBLUP*:* $\widehat{\bar{\boldsymbol{y}}}_{ds}^{MEBLUP}$ for $\rho_e = -0.7,\ \rho_u = 0.3$ and $\rho_e = 0.7, \rho_u = 0.3$ cases.
6. Estimate the vector of MSEs of $\widehat{\bar{\boldsymbol{y}}}_{ds}^{MEBLUP}$ on each sample *s* via the parametric bootstrap proposed in section 3 with $B = 500$ replications.

The results are evaluated via the empirical MSE (EMSE), which is considered to be the "truth", the bootstrap MSE across the $S = 1{,}000$ simulations, and the relative bias (RBIAS) for each small area *d*. These quantities are respectively defined by the following estimators:

$$\boldsymbol{EMSE}(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP}) = S^{-1}\sum_{s=1}^{S}(\widehat{\bar{\boldsymbol{y}}}_{ds}^{MEBLUP} - \bar{\boldsymbol{y}}_{ds})(\widehat{\bar{\boldsymbol{y}}}_{ds}^{MEBLUP} - \bar{\boldsymbol{y}}_{ds})', \tag{12}$$

$$\boldsymbol{mse}_{*}^{B}(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP*}) = S^{-1}\sum_{s=1}^{S}\boldsymbol{mse}_{*s}(\widehat{\bar{\boldsymbol{y}}}_{ds}^{MEBLUP*}), \tag{13}$$

where we denote the bootstrap MSE of sample *s* in area *d* by: $\boldsymbol{mse}_{*s}(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP*})$

$$\boldsymbol{RBIAS}\left[\boldsymbol{mse}_{*}\left(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP*}\right)\right] = S^{-1}\sum_{s=1}^{S}\left[\boldsymbol{mse}_{*s}\left(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP*}\right) - \boldsymbol{EMSE}\left(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP}\right)\right]/\boldsymbol{EMSE}\left(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP}\right). \tag{14}$$

$\boldsymbol{EMSE}(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP})$ denotes the empirical mean squared error of $\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP}$, where $\bar{\boldsymbol{y}}_{ds} = N_d^{-1}\sum_{i=1}^{N_d}\boldsymbol{y}_{dis}$. $\boldsymbol{mse}_{*}(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP*})$ denotes the average of the bootstrap MSEs (based on $B = 500$ replicates) across the $S = 1{,}000$ samples drawn in the simulation, and $\boldsymbol{RBIAS}[\boldsymbol{mse}_{*}(\widehat{\bar{\boldsymbol{y}}}_{d}^{MEBLUP*})]$ denotes its relative bias.

The same estimators can be written for the univariate case both for the bootstrap and Prasad-Rao (PR) approximations. We do not review the Prasad-Rao analytical approximation in this

paper, thus we refer to Prasad and Rao (1999) for theoretical details. The reader may want to refer to González-Manteiga et al. (2008a) for the parametric bootstrap for the univariate case.

4.2 Results

Here we compare first the MSE estimates obtained via the Prasad-Rao analytical approximation with the MSE estimates obtained by parametric bootstrap for the univariate case. Table 1 shows the descriptive statistics and bias of the Prasad-Rao and bootstrap estimators for univariate EBLUP across the small areas. It can be seen that the Prasad-Rao MSEs analytical approximations are slightly more biased than the bootstrap MSEs (by comparing the EMSE with its mean across the $S = 1{,}000$ samples) under our scenario. Figure 1 and Figure 2 show the relative bias of the MSEs; these show that the Prasad-Rao MSE approximation slightly overestimates the true MSE for some areas. This is particularly true for Y_1. For more details on the Prasad-Rao approximation compared to the bootstrap for univariate EBLUP via simulation studies we refer to González-Manteiga et al. (2008a).

Insert Table 1 about here

Insert Figure 1 about here

Insert Figure 2 about here

In Table 2 we compare the results of the univariate with the multivariate bootstrap MSE estimation in terms of reduction in MSE and bias. We calculate the relative percentages of reduction in terms of EMSE (and bootstrap MSE) as follows: $\Delta_{dk} =$

$\frac{EMSE(\hat{\bar{Y}}_{dk}^{MEBLUP})-EMSE(\hat{\bar{Y}}_{dk}^{EBLUP})}{EMSE(\hat{\bar{Y}}_{dk}^{EBLUP})} \cdot 100$, for $d = 1, \dots, D$, where $k = 1,2$ denotes the index of the k^{th} component of the MEBLUP means vector or the k^{th} variable in case of univariate EBLUP. These are shown in parentheses ( ). We also show the median across the small areas of the following quantities: empirical MSE ($EMSE$), bootstrap MSE estimates (mse_*) and relative bias % ($RBIAS(mse_*)\%$), and relative percentages of reduction in terms of EMSE (and bootstrap MSE) (Δ_d). We provide the median across the small area as a robust central tendency index to avoid the impact of extreme values in some small areas (Giusti et al., 2013; Chambers et al., 2011). Figure 3 and Figure 4 show the comparisons of the bootstrap MSEs estimated for the EBLUPs and MEBLUPs for the opposite signs case only. It can be seen that, in line with the EMSEs, the multivariate bootstrap procedure provides predictions with lower variability than the univariate approach, and the MSE estimates show no noticeable bias across the small areas. When the population size in area *d,* N_d*,* increases, Δ_{dk} becomes smaller. The percentages of reduction in terms of MSE are smaller in the case of same signs of the correlation coefficients in the variance-covariance matrices.

Insert Table 2 about here

Insert Figure 3 about here

Insert Figure 4 about here

4.3 Final remarks on the simulation study

The percentage of reductions in terms of MSE (and EMSE) may depend on the magnitude and sign of ρ_e and ρ_u as well as the intra-class correlation coefficient. As Datta et al. (1999) points out that when ρ_e and ρ_u have opposite signs, the multivariate model performs better in terms of MSE than the univariate modelling case. It can be seen that when the signs in the variance-covariance matrices are the same the percentages of reduction in terms of MSE of the multivariate EBLUP over the univariate ones are smaller than in the case of opposite signs.

Our bootstrap procedure performs well under the model assumptions, and we can see appreciable gains in efficiency in terms of MSE over the univariate modelling. Also, we note that there is no bias in the estimates of the MSE.

5. Application to Corn and Soy Bean Data

We apply our multivariate bootstrap method to the well-known corn and soy bean data of the LANDSAT data that was used in Battese et al. (1988) comparing the multivariate and univariate models. LANDSAT comprises survey and satellite data for corn and soy beans for 12 Iowa counties, obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture and from land observatory satellites during the 1978 growing season. The data file consists of $n = 37$ observations, $D = 12$ areas, and the following variables:

- CornHec: hectares of corn (Y_1);
- SoyBeansHec: hectares of soy beans (Y_2);
- CornPix: number of pixels of corn in sample segment within county (x_1);
- SoyBeansPix: number of pixels of soy beans in sample segment within county (x_2).

As shown, the county means of number of pixels per segment of corn and soy beans, from satellite data, for 12 counties in Iowa are also used where we have the population size, sample size, and means of these auxiliary variables. These data files can be downloaded from Molina and Marhuenda (2015). In order to provide better modeling fit we applied Box-Cox family transformations (Box and Cox, 1964) to the response variables.

Insert Figure 5 about here

Insert Figure 6 about here

Figure 5 and Figure 6 show the RMSE of the univariate and multivariate EBLUPs where the small areas are ordered by growing sample sizes. It can be seen that the multivariate bootstrap algorithm provides estimates with smaller variability than the univariate case as was confirmed in the simulation study. The model diagnostics show good model fitting in both cases.

6. Conclusion

In this paper we proposed the use of parametric bootstrap for estimating MSEs for vectors of means of small domains under a multivariate mixed effects model for unit-level SAE. The multivariate SAE is more appropriate than the univariate SAE in the case of correlated responses. Indeed, in this case, multivariate mixed effects models may lead to more reliable estimates than the univariate BHF model. This, of course, needs to be taken into account when estimating the MSE and hence we have proposed the parametric bootstrap for the MEBLUP. In the simulation study we assessed empirically the behaviour of our approach for

estimating the MSE and in particular the bias. Our results are in line with the literature and no bias is shown.

Although this paper focuses on vectors of means as the target inferential parameter, this bootstrap procedure can be extended to other quantities in a multivariate setting. Non-parametric bootstrap procedures could be studied in future work, and comparisons between the two methodologies would be useful for practitioners. Normality assumptions can be relaxed according to Hall and Maiti (2006). Furthermore, hybrid bootstrap MSE estimators should be considered. González-Manteiga et al. (2008a) studied hybrid bootstrap MSE estimators which are second-order unbiased. Other interesting extensions to this paper may involve the study of robustness to non-normality.

Acknowledgments

This research was financially supported by the UK Economic and Social Research Council (ESRC), grant number ES/J500094/1.

Appendix: Specification of the R functions used

Here we describe the main R packages (and functions) that we used to conduct the simulation study and application.

1. Estimation of small area means and bootstrap MSE under the univariate BHF model: 'sae' R package (Molina and Marhuenda, 2015):
   - Required packages: nlme, MASS,
   - Functions: eblupBHF( ) and pbmseBHF( ).
2. Estimation of the multivariate mixed effects model parameters ($\boldsymbol{\Sigma}_e, \boldsymbol{\Sigma}_u, \boldsymbol{\beta}$) described in (3): 'mlmmm' R package (Yucel, 2010):
   - Function: mlmmm.em.

References

Baillo, A. and Molina, I. (2009). Mean Squared Errors of Small-Area Estimators Under a Unit-Level Multivariate Model. Statistics 43:553-569.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An Error-Components model for Prediction of County crop areas using Survey and Satellite data. Journal of the American Statistical Association 83(401): 28-36.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. Journal of Royal Statistical Society Series B 26:211-246.

Chambers, R., Chandra, H., and Tzavidis (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. Survey Methodology 37:153-170.

Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean Squared Error of Empirical Predictor. Annals of Statistics 32:818-840.

Datta, G. S., Day, B. and Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. Journal of Statistical Planning and Inference 75:269-279.

Datta, G.S. and Lahiri, P. (2000). A Unified Measure of Uncertainty pf Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems. Statistica Sinica 10:613-627.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

Fuller, W. A. and Harter, R. M. (1987). The Multivariate Components of Variance Model for Small Area Estimation, in R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh (Eds.), *Small Area Statistics*, New York: Wiley, 103-123.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008a). Bootstrap Mean Squared Error of a Small-Area EBLUP. Journal of Statistical Computation and Simulation 78:443-462.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008b). Analytic and bootstrap approximations of prediction errors under multivariate Fay-Herriot model. Computational Statistics and Data Analysis 52:5242-5252.

Giusti, C., Tzavidis, N., Pratesi, M. And Salvati, N. (2013) Resistance to Outliers pf M-Quantile and Robust Random Effects Small Area Models. Communications in Statistics

– Simulation and Computation 43:549-568.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer.

Hall, P. and Maiti, T. (2006). Nonparametric Estimation of Mean-Sqaured Prediction Error in Nested-Error Regression Models. Annals of Statistics 34:1733-1750.

Kackar, R.N. and Harville, D.A. (1981). Unbiasedness of Two-Stage Estimation and prediction Procedure for Mixed Linear Models. Journal of the American Statistical Association 79:853-862.

Marchetti, S., Tzavidis, N. and Pratesi, M. (2012). Non-parametric Bootstrap Mean Squared Error Estimation for M-quantile Estimators of Small Area Averages, Quantiles and Poverty Indicators. Computational Statistics and Data analysis 56:2889-2902.

Mardia, K. V., and Marshall, R. J. (1984). Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression. Biometrika 71(1).

Molina, I. (2009). Uncertainty under a multivariate nested-error regression model with logarithmic transformation. Journal of Multivariate Analysis 100:963-980.

Molina, I., and Marhuenda, Y. (2015). sae: An R Package for Small Area Estimation. The R Journal 7(1):81-98.

Prasad, N.G.N. and Rao, J.N.K. (1990). The Estimation of Mean Squared Error of Small-Area Estimators. Journal of the American Statistical Association 85:163-171.

Rao, J. N. K. and Molina, I. (2015). *Small area estimation*. New York: Wiley.

Schafer, J. L., and Yucel, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. Journal of Computational and Graphical Statistics 11:437-457.

Sweeting, T. J. (1980). Uniform Asymptotic Normality of the Maximum Likelihood Estimator. The Annals of Statistics 8.

Yucel, R. (2010) mlmmm: ML estimation under multivariate linear mixed models with missing values. R package version 0.3-1.2. Retrieved from http://CRAN.R-project.org/package=mlmmm.

Tables and Figures

Estimator	Mean	Median	IQR	SD	RBIAS%
$mse^{PR}\left(\widehat{\bar{Y}}_1^{EBLUP}\right)$	18.42	18.04	7.48	4.29	2.32%
$mse_*\left(\widehat{\bar{Y}}_1^{EBLUP}\right)$	18.29	17.90	7.14	4.26	-0.20%
$mse^{PR}\left(\widehat{\bar{Y}}_2^{EBLUP}\right)$	14.03	13.34	7.28	4.31	6.31%
$mse_*\left(\widehat{\bar{Y}}_2^{EBLUP}\right)$	14.09	13.41	7.27	4.34	3.45%

Table 1 Descriptive statistics and relative bias of the Prasad-Rao and bootstrap estimators for univariate EBLUP MSE across small areas, $EMSE(\widehat{\bar{Y}}_1^{EBLUP}) = 17.98$, $EMSE(\widehat{\bar{Y}}_2^{EBLUP}) = 12.75$, for $\rho_e = -0.7$ and $\rho_u = 0.3$.

Correlation structure	Performance measure	EBLUP		MEBLUP	
		$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_1$	$\bar{Y}_2$
$\rho_e = -0.7$,	$EMSE$	17.98	12.75	16.26 (-9.25)	10.62 (-17.63)
$\rho_u = 0.3$	mse_*	17.90	13.41	16.34 (-9.40)	11.12 (-17.48)
	$RBIAS(mse_*)\%$	-0.20%	3.45%	-0.06%	1.50%
$\rho_e = 0.7$,	$EMSE$	17.27	12.64	17.43 (1.10%)	12.06 (-6.25%)
$\rho_u = 0.3$	mse_*	18.07	13.38	17.88 (-1.41)	12.21 (-.9.00%)
	$RBIAS(mse_*)\%$	3.56%	6.41%	1.72%	2.55%

Table 2 Empirical mean squared error, bootstrap MSE, relative bias median results across the small areas: EBLUP and MEBLUP estimates – parametric bootstrap. (Δ_k shown in parenthesis).

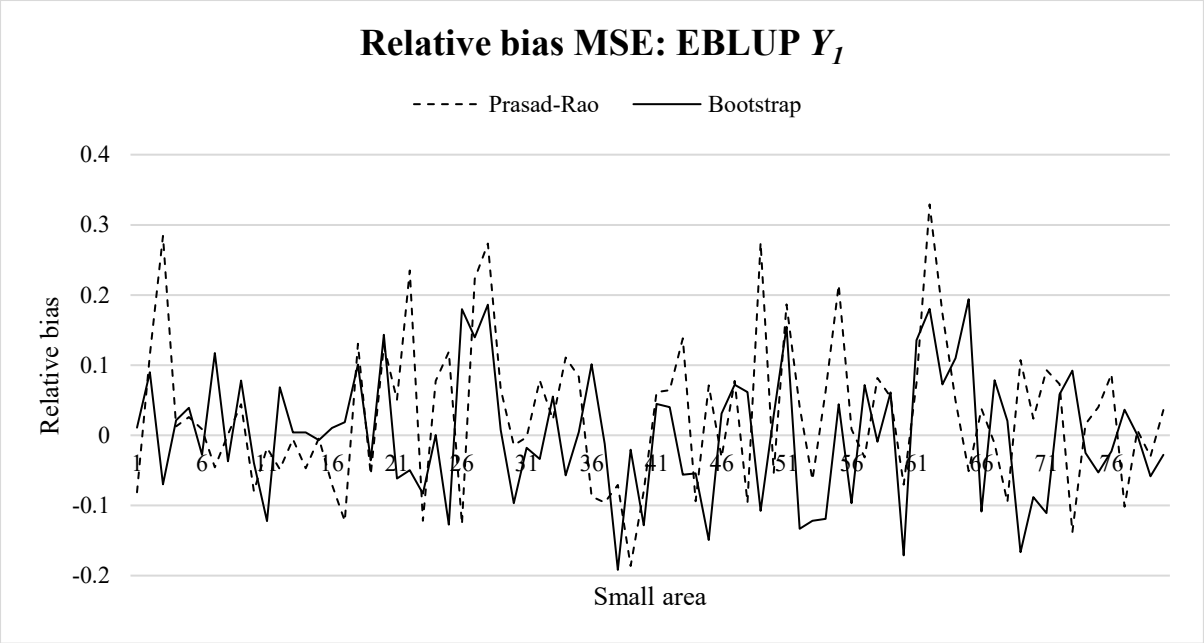

Figure 1 Relative bias of univariate EBLUPs' MSEs of Y_1 estimated via Prasad-Rao approximation and parametric bootstrap for $\rho_e = -0.7$ and $\rho_u = 0.3$, ordered by increasing N_d.

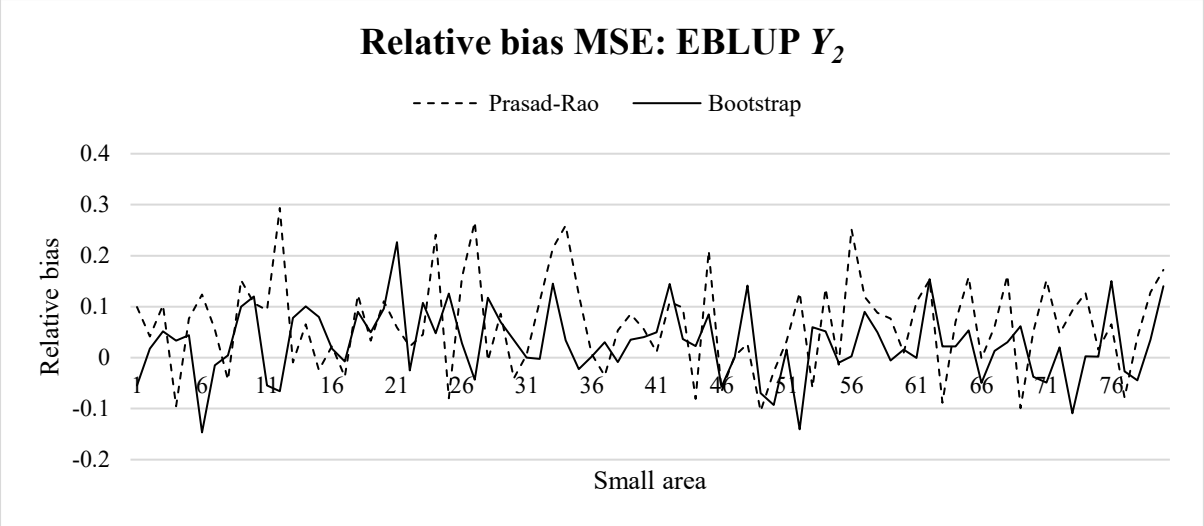

Figure 2 Relative bias of univariate EBLUPs' MSEs of Y_2 estimated via Prasad-Rao approximation and parametric bootstrap for $\rho_e = -0.7$ and $\rho_u = 0.3$, ordered by increasing N_d.

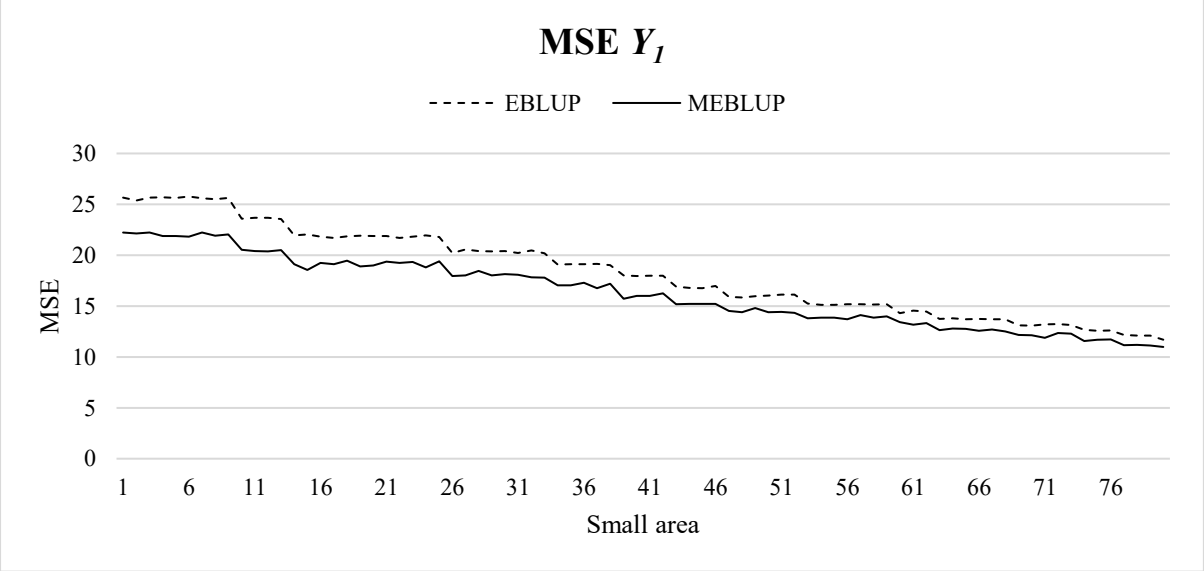

Figure 3 Bootstrap MSEs Y_1: comparison between EBLUP and MEBLUP $\rho_e = -0.7$ and $\rho_u = 0.3$, ordered by increasing N_d.

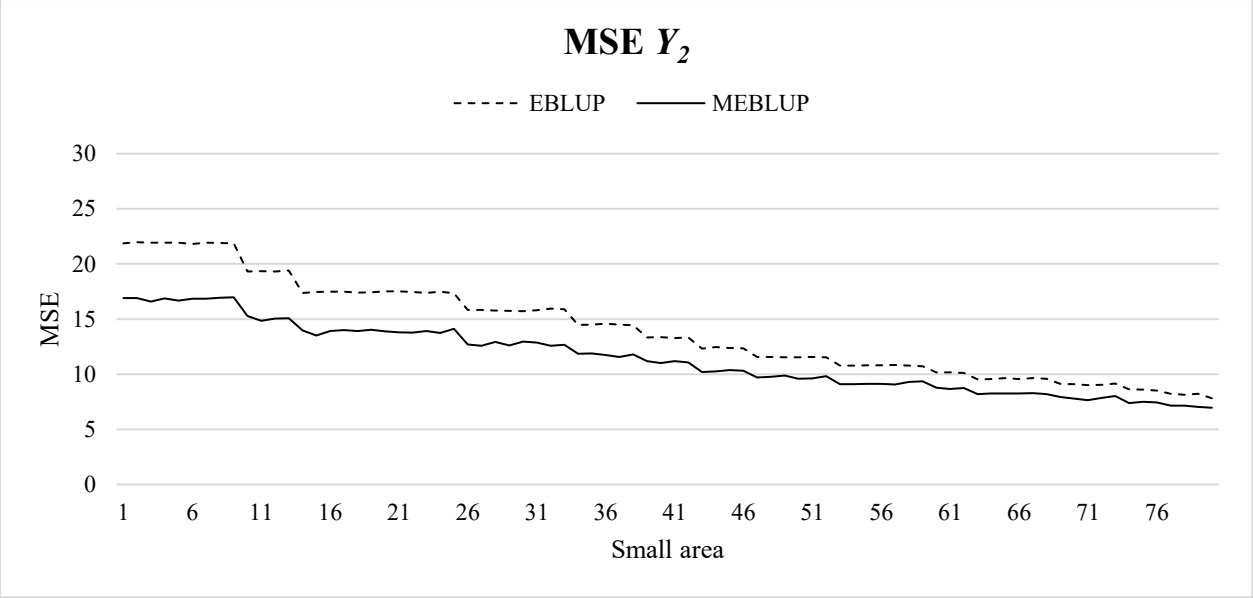

Figure 4 Bootstrap MSEs Y_2: comparison between EBLUP and MEBLUP $\rho_e = -0.7$ and $\rho_u = 0.3$ ordered by increasing N_d.

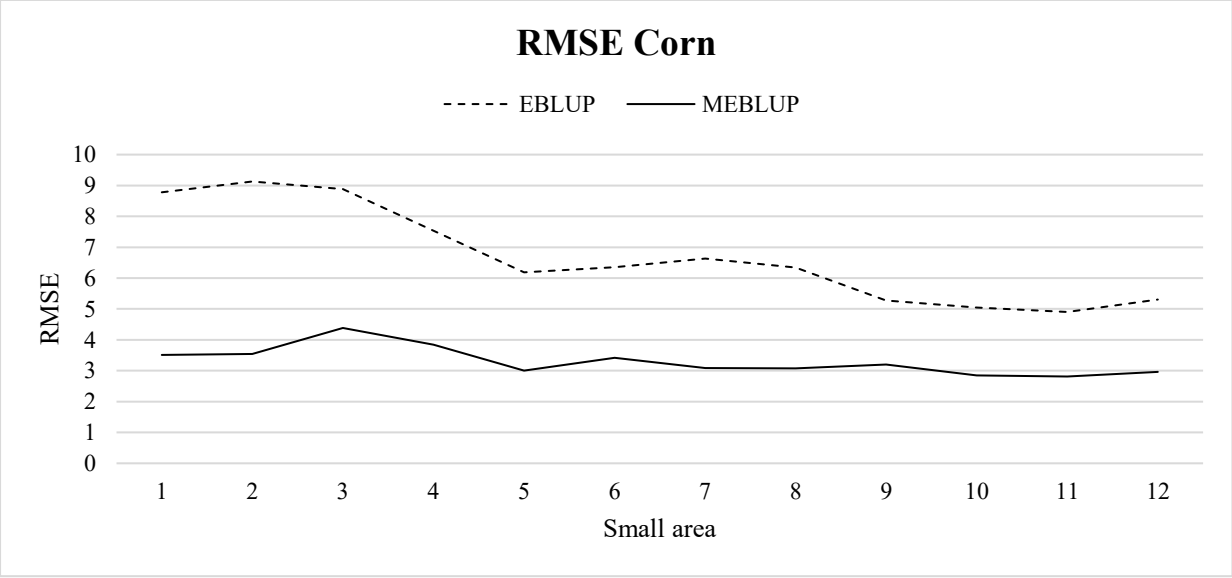

Figure 5 Bootstrap RMSEs corn: comparison between EBLUP (---) and MEBLUP (__).

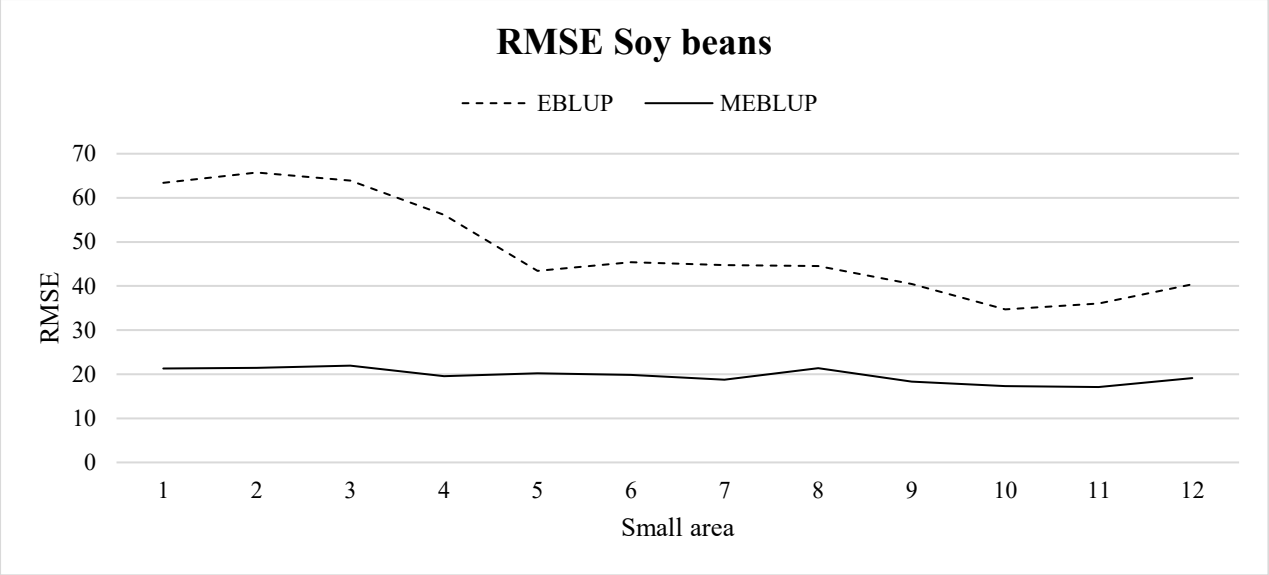

Figure 6 Bootstrap RMSEs soy beans: comparison between EBLUP (---) and MEBLUP (__).